Twitter activity of the US members of Congress from 2007 to 2017

Author: Guus Bouwens
Date: 04-08-2022

Social media communication patterns ¶

The main analysis aims to get insights in two major topics.

- Some key points about Congress members Twitter presence and how do they perform and interact.
- Common patterns about what Congress members communicate.

To address these issues three specific questions were selected to obtain some of the information sought.

1. The relationship between popularity and activity on Twitter as well as the states population of members of Congress.
    - Do members of Congress from more populous states have more followers?
    - Do members of Congress who post the most have the most followers?
2. Do the colors of the images of the Twitter accounts (profile and cover) are mainly split between red and blue ?
3. What are the most used words/hashtags by members of Congress?

Hypotheses

1. The more a member of Congress tweets the more he/she is being followed. The same assumption is made regarding the population of the states.
    - *However, the date of creation of the twitter account should strongly model these assumptions.*
2. There is a good chance that the colors red and blue are used a lot depending on the parties of members of Congress.
    - *However, the result could be biased by how to measure/calculate the dominant color of the whole image or a piece of it.*
3. Getting strong predictions of which words/hashtags are most used is complex without any knowledge, but chances are that among them are the words Democrat and Republican, as well as Obama and Trump (and/or a compound word like Obamacare).

Analysis approach

After collecting, cleaning and transforming the data some usual steps are processed.

1. Getting features descriptive statistics: range, outstanding data, distribution.
2. Quantitative analysis: relationships measurements between variables, regression.
3. Qualitative analysis: in this case it's about text analysis.

Data architecture

Two JSON files make up the dataset, one gathering the Twitter accounts of members of Congress and the other all the actions (tweets, retweets, etc.) of these accounts. From the features needed to answers our problematics, the following EDR can summary the tables processed.
US states and their population were gathered and add in a table as well.

Image

Data analysis

In [1]:

```
##
import pandas as pd
import numpy as np
import math
from scipy import stats
from IPython.display import display

import nltk
from nltk import word_tokenize
from nltk.corpus import stopwords
from nltk.probability import FreqDist

import datetime
from datetime import datetime as dt
import re
import time

import sqlite3
from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())

import os

import plotly.express as px
import matplotlib
import matplotlib.pyplot as plt
import matplotlib.pylab as plb
from matplotlib import cm
%matplotlib inline

pd.set_option('display.max_rows', 11)
pd.set_option('display.max_colwidth', None)
pd.set_option('display.max_columns', None)


def fivenum(table_name, table_col):

    mean = pysqldf(f"""SELECT
                            "MEAN" AS Stat,
                            CAST(AVG({table_col}) AS int) AS {table_col}
                        FROM {table_name}""")

    median = pysqldf(f"""SELECT
                            "MEDIAN" AS Stat,
                            {table_col} AS {table_col}
                        FROM {table_name}
                        ORDER BY {table_col}
                        LIMIT 1
                        OFFSET (SELECT COUNT(*) FROM {table_name}) / 2""")

    min = pysqldf(f"""SELECT
                            "MIN" AS Stat,
                            MIN({table_col}) AS {table_col}
                        FROM {table_name}""")

    max = pysqldf(f"""SELECT
                            "MAX" AS Stat,
                            MAX({table_col}) AS {table_col}
                        FROM {table_name}""")
    return pd.concat([min, median, mean, max])

%load_ext sql
%sql sqlite:///data/02_processed/CongressTweets.db

users_tbl = %sql SELECT * FROM users
users_tbl = users_tbl.DataFrame()

tweets = %sql SELECT created_at, screen_name, text FROM tweets
tweets = tweets.DataFrame()
```

```
 * sqlite:///data/02_processed/CongressTweets.db
Done.
 * sqlite:///data/02_processed/CongressTweets.db
Done.
```

Initial insights: Twitter popularity and activity

Among the members of Congress on Twitter, there are extreme cases who have a very large number of followers or write a very large number of tweets and far ahead of their colleagues.

image1

Initial insights: Twitter profile colors

From the observation of the distributions of the different variables linked to the colors, it is difficult to obtain convincing results. However, among the 5 color features metrics, `profile link color` seems to contain mostly only shades of red and blue. There is therefore at least one of the variables which verifies our initial hypothesis.

- There are several shades for the same color (or range), which dissociates the data instead of grouping them and consequently makes the analysis much more difficult. Therefore, this analysis is subsequently dropped.

image1

Initial insights: tweets text

In order to get an overview, a variable that counts the individual words according to NLP rules was created.

Most of tweets follows a similar pattern with a number of words per tweet near the average value of all Congress members tweets. This is expected as Twitter limit the number of characters by tweet to a low amount, 140 in total. This necessarily implies having a limited number of words but leaves a little flexibility in the way of constructing a concise message.

image1

Relationships analysis: Twitter popularity and activity

Initial variables weren't under normal law assumptions, as seen previously with their distribution plots. A logarithmic transformation was used to create suitable metrics for the analysis.

There are clearly correlations (statisticly significant) and linear responses between number of followers and number of tweets, as well as number of followers and the account creation date.

- Every 35% increase of written tweets leads in average a 24% increase of followers.
- Each year passed leads in average a 14% increase of followers.

However, the number of inhabitants of the states represented by members of Congress has no impact on the number of followers. Which is the reverse of the initial hypothesis.

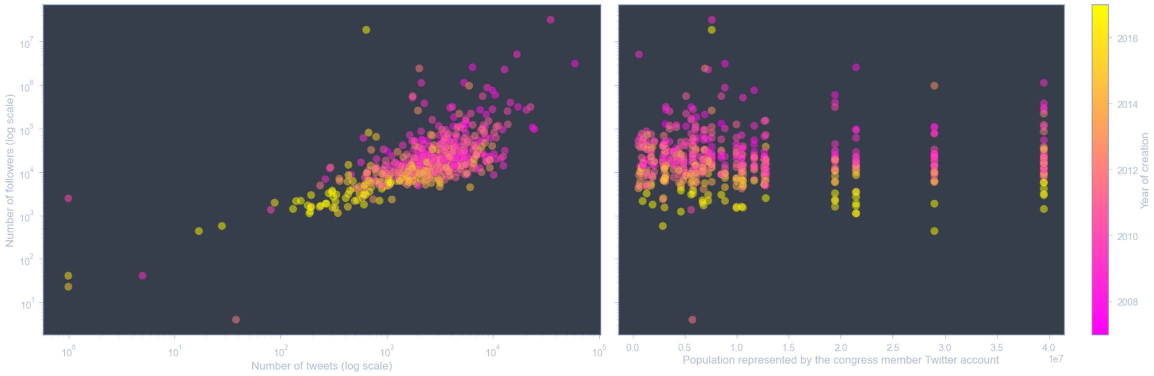
In [53]:

```
##
conn = sqlite3.connect('data/02_processed/CongressTweets.db')

with conn:
    c = conn.cursor()
    query = ("""SELECT
                    created_at,
                    followers_count AS followers,
                    name            AS pseudo,
                    screen_name     AS account_name,
                    statuses_count  AS tweets_count,
                    users.state     AS state,
                    population,
                    percentage      AS pop_percent
                FROM
                    users
                LEFT JOIN
                    states
                ON
                    users.state = states.state
                """)
    congress_members = pd.read_sql_query(query,conn)

conn.close()

congress_members['date_creation'] =  pd.to_datetime(congress_members['created_at']).dt.
year

##
def scatter(congress_members):
    fig, ax = plt.subplots(1,2, sharey=True)
    fig.set_size_inches(25,8)

    x = (congress_members['tweets_count']+1)
    y = congress_members['followers']
    z = congress_members['date_creation']

    ax[0].scatter(x, y,
                s = 150, c=z, cmap='spring', alpha=0.5,
                edgecolors='black')

    ax[0].set_xscale('log')
    ax[0].set_yscale('log')

    ax[0].set_xlabel('Number of tweets (log scale)')
    ax[0].set_ylabel('Number of followers (log scale)')

    ##

    x = congress_members['population']
    y = congress_members['followers']
    z = congress_members['date_creation']

    ax[1].scatter(x, y,
                s = 150, c=z, cmap='spring', alpha=0.5,
                edgecolors='black')

    # Cmap legend
#     fig.cbar
```

```
    cbar = fig.colorbar(cm.ScalarMappable(norm=matplotlib.colors.Normalize(vmin=min(z),
vmax=max(z), clip=True),
                                          cmap='spring'),
                        ax=ax[1])
    cbar.set_label('Year of creation')

    ax[1].set_yscale('log')

    ax[1].set_xlabel('Population represented by the congress member Twitter account')
    plt.tight_layout()

    plt.show()
```

In [54]:

```
scatter(congress_members)
```



Text analysis: tweets patterns

After removing meaningless words (i.e. stop words like "the, "a", etc ...), the motion char below show the top 10 most written words between 2008 and 2017.

In [2]:

```
##
tweets = %sql SELECT created_at, screen_name, text FROM tweets
tweets = tweets.DataFrame()

# Data cleaning: stop words, lowercase
stopwrds_lst = list(stopwords.words('english'))
stopwrds_lst.extend(('w', "'", 'th', "'s", "'t", "n't", "http", "https", "'m"))
stopwrds_lst = set(stopwrds_lst)

tweets['clean_text'] = tweets['text'].str.lower()
tweets['clean_text'] = tweets['clean_text'].str.replace('[&@#_$+,:;=?~{}|…"`’\\<>.^*°
()%!\\-\\[\\]/0-9]|(&amp)|(amp)', "")
tweets['clean_text'] = tweets['clean_text'].str.replace('thanks', "thank")
tweets['clean_text'] = tweets['clean_text'].str.replace('americans', "american")
tweets['clean_text'] = tweets['clean_text'].transform(word_tokenize)
tweets['clean_text'] = tweets['clean_text'].apply(lambda x: [words for words in x if wo
rds not in stopwrds_lst])
```

```
 * sqlite:///data/02_processed/CongressTweets.db
Done.

<ipython-input-2-d9137f6eee01>:11: FutureWarning: The default value of reg
ex will change from True to False in a future version.
  tweets['clean_text'] = tweets['clean_text'].str.replace('[&@#_$+,:;=?~{}
|…"`’\\<>.^*°()%!\\-\\[\\]/0-9]|(&amp)|(amp)', "")
```

In [6]:

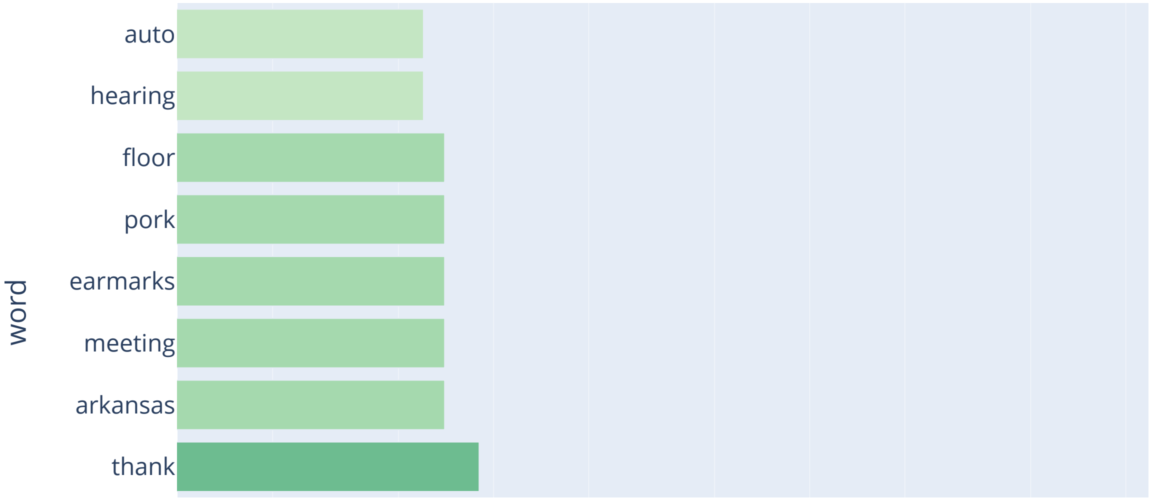
```
# Get top 10 words for each year
tweets['created_at'] = tweets['created_at'].apply(lambda x: x[0:4])

tidy_df = pd.DataFrame()
slice_df = pd.DataFrame()

for year in tweets['created_at'].unique():

    filt = tweets['created_at'] == year
    agg_corpus = ' '.join(tweets.loc[filt, 'clean_text'].str.join(' ')).split(' ') #mus
t be tokens single list

    fdist = FreqDist(agg_corpus)
    top10_words = fdist.most_common(10)

    slice_df['word'] = [item[0] for item in top10_words]
    slice_df['frequency'] = [item[1] for item in top10_words]
    slice_df['year'] = year

    tidy_df = tidy_df.append(slice_df)
```

In [7]:

```
# Create a motion chart
fig = px.bar(tidy_df,
       x="frequency", y="word", animation_frame="year",
       range_x=[1,10e4], log_x=True,
       color="frequency", color_continuous_scale=px.colors.sequential.Blugrn)
```

In [8]:

```
fig
```



1. Many of the most used words are hardly surprising either by the lexical field used or by the political context. Indeed, "american", "US", "bill", "care", "health", "house", "congress" are all words that are strongly related to the nation, to the function, and to the law.
2. Clearly one would have expected that one of the most used words on Twitter would be "rt", but this was not taken into account when making assumptions.
3. The words "obamacare" and "trump" appeared in the top 10 most quoted words in 2013 and 2017 respectively. The initial hypothesis is therefore partially verified.

Hypotheses conclusions

1. Twitter popularity and activity assumptions:
   - The more the number of tweets, the more the number of followers. ✅
   - The older a Twitter account is, the more followers it has. ✅
   - Population of the state represented by the congress member Twitter account have not any impact on the number of followers. ❌
2. Twitter profile color assumptions:
   - Data and their metrics didn't allow clear conclusions. ❕
3. Tweets common patterns assumptions:
   - "Democrat" and "Republican" words didn't appeared. ❌
   - "Obama" and "Trump" related words appeared some years among the top 10 most used words. ✅

Further potential analysis

To get additional informations regarding the current conclusions, here's some ideas to consider:

- Getting Congress members party could lead to hidden patterns among the previous steps taken in this report. But it requires a data collection substantial work.
- Finding a way to group shades of colors and continue the analysis initiated.
- Doing the famous sentimental text analysis and linking insights with the other features.